SelfLRE: Self-refining Representation Learning for Low-resource Relation Extraction

Xuming Hu
Tsinghua University
hxm19@mails.tsinghua.edu.cn

Junzhe Chen
Tsinghua University
chenjz20@mails.tsinghua.edu.cn

Shiao Meng
Tsinghua University
msa21@mails.tsinghua.edu.cn

Lijie Wen
Tsinghua University
wenlj@tsinghua.edu.cn

Philip S. Yu
University of Illinois at Chicago
psyu@cs.uic.edu

ABSTRACT

Low-resource relation extraction (LRE) aims to extract potential relations from limited labeled corpus to handle the problem of scarcity of human annotations. Previous works mainly consist of two categories of methods: (1) Self-training methods, which improve themselves through the models' predictions, thus suffering from confirmation bias when the predictions are wrong. (2) Self-ensembling methods, which learn task-agnostic representations, therefore, generally do not work well for specific tasks. In our work, we propose a novel LRE architecture named SelfLRE, which leverages two complementary modules, one module uses self-training to obtain pseudo-labels for unlabeled data, and the other module uses self-ensembling learning to obtain the task-agnostic representations, and leverages the existing pseudo-labels to refine the better task-specific representations on unlabeled data. The two models are jointly trained through multi-task learning to iteratively improve the effect of LRE task. Experiments on three public datasets show that SelfLRE achieves 1.81% performance gain over the SOTA baseline. Source code is available at: https://github.com/THU-BPM/SelfLRE.

CCS CONCEPTS

• **Computing methodologies** → **Information Extraction**.

KEYWORDS

Low-resource Relation Extraction, Representation Learning

ACM Reference Format:
Xuming Hu, Junzhe Chen, Shiao Meng, Lijie Wen, and Philip S. Yu. 2023. SelfLRE: Self-refining Representation Learning for Low-resource Relation Extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592058

1 INTRODUCTION

Relation Extraction (RE) aims to extract relation between entities from corpus and obtain triplets: `{Owl, Component-Whole, Claw}`

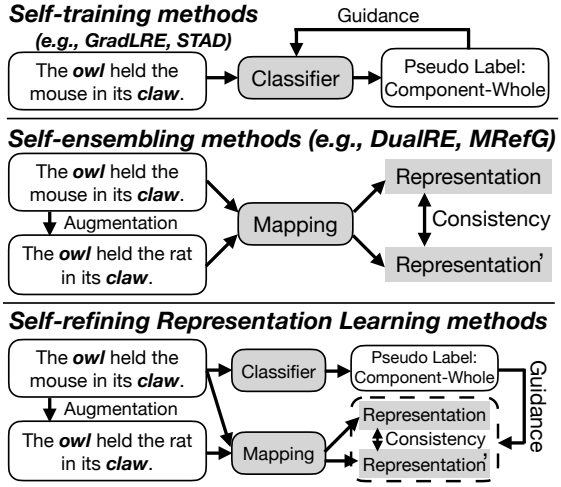

Figure 1: Comparison of self-training methods, self-ensembling methods, and our proposed self-refining representation learning methods. For our methods, the pseudo-labels generated by self-training methods could refine task-specific representations on self-ensembling methods. The refined representations enable more correct pseudo-labels by optimizing classifiers for self-training methods.

(In Figure 1) for downstream information retrieval (IR) tasks such as web searching [1, 4, 11] and question answering [2, 6, 20]. The current relation extraction networks need to rely on large amounts of high-quality labeled data to achieve decent performance. However, it would be labor-intensive to obtain labeled data. Therefore, the low-resource relation extraction (LRE) task is crucial to improve the ability of the model by utilizing unlabeled data [3, 27]. As shown in Figure 1, existing methods adopt two types of methods: (1) Self-training methods and (2) Self-ensembling methods to utilize the unlabeled data. Self-training methods (e.g., Co-training [28], GradLRE [9], and STAD [27]) leverage the fine-tuned models to pseudo-label the unlabeled data, and adopt the pseudo-labeled data as the guidance to continue to optimize the model. However, these methods inevitably suffer from the confirmation bias when the pseudo labels are wrong. As incorrect pseudo-labeled data is continuously added to the labeled data for iterative training, the model will drift away from the local optimum. Self-ensembling methods (e.g., Mean Teacher [24], DualRE [17], and MRefG [16]) first adopt data augmentation methods to generate sentences with similar relational semantics, and leverages the fine-tuned mapping model to obtain representations of two sentences. Inspired by contrastive

SIGIR '23July 23–27, 2023Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592058

learning methods [10, 13, 18], the similar relational representations will pull closer, and vice versa. However, these methods can only learn task-agnostic representations, while RE task-specific representations, such as relation labels, cannot be learned specifically. *Why not combine the strengths of self-training and self-ensembling methods while avoiding the shortcomings?*

In this paper, we propose a novel self-refining representation learning architecture for LRE task named SelfLRE, which treats self-training and self-ensembling methods as two complementary modules. As shown in Figure 1, we first adopt the fine-tuned model to generate pseudo labels on unlabeled data. Then we use synonym replacement to obtain an augmented sentence with the same relational semantics as the original sentence, and adopt the fine-tuned mapping network to obtain the corresponding sentence representations. Certainly, these semantic representations are task-agnostic. Therefore, we solicit the pseudo-labels generated by the classifier to pull the representations under the same pseudo-relational label to be close to each other, so as to refine the relational representations specific to the RE task. Thanks to two complementary methods, the pseudo labels generated by self-training methods could be corrected by whether the semantic representations of the same relation labels are close. The representations generated by the self-ensembling methods could obtain task-specific guidance through pseudo-labels, thereby pulling closer representations under the same pseudo-labels. To summarize, the main contributions of this work are as follows: (1) We propose a novel self-refining representation learning architecture named SelfLRE for LRE task, which treats self-training and self-ensembling methods as two complementary modules. Pseudo-labels can refine task-specific representations. Task-specific representations enable more correct pseudo-labels by optimizing classifiers. (2) Experiments on three public datasets show that SelfLRE achieves 2.68% performance gain over the SOTA baseline. Extensive analysis validates the effectiveness of SelfLRE.

2 TASK FORMULATION

Our task involves labeled and unlabeled sets for low-resource RE setting. For labeled data: $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N_l}$, where y_i are one-hot ground truth labels and N_l is the size of labeled samples. For unlabeled data $\mathcal{U} = \{(u_i)\}_{i=1}^{N_u}$, with N_u being the size of unlabeled samples. To create a contextual representation of an input sentence x, we use BERT as a text encoder. The model includes a classifier head $fc(\cdot)$ that produces a probability distribution of sentence x over different classes $p(y|x) = fc(BERT(x))$, and a mapping head $map(\cdot)$ that maps the contextual representation obtained from BERT [5] to a regularized low-dimensional embedding $\boldsymbol{e}(x) = map(BERT(x))$. The two modules leverage unlabeled data to complement each other and the fine-tuned BERT with classifier head $fc(BERT(x))$ will be evaluated as the final model.

3 MODEL

3.1 Pseudo Label Generation

The Pseudo Label Generation aims to obtain pseudo-labels for unlabeled data. We begin by fine-tuning a *BERT* model using labeled data $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N_l}$, where y_i are one-hot ground truth labels. For a sentence $x = [t_1, t_2, \cdots, t_T]$ with corresponding entities $E1$ and $E2$, we follow Soares et al. [22] by including four special tokens

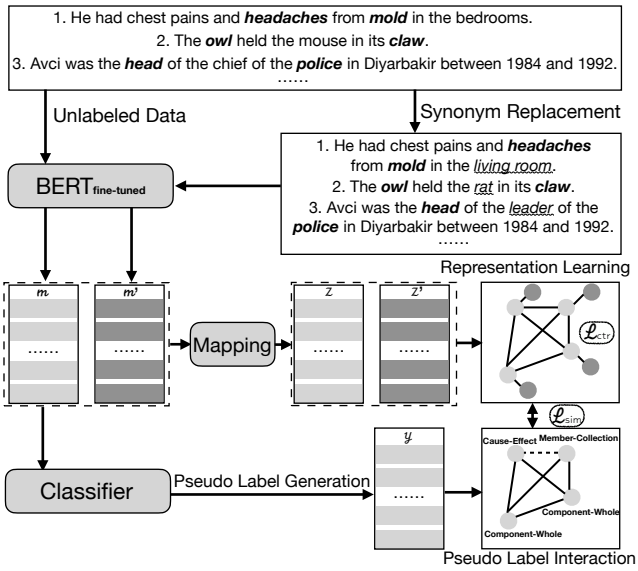

Figure 2: Architecture of SelfLRE.

to denote the start and end of $E1$ and $E2$. We introduce the $[E1_{start}]$, $[E1_{end}]$,$[E2_{start}]$, $[E2_{end}]$ and inject them to x:

$$\begin{aligned} x = &[t_1, \cdots, [E1_{start}], t_i, \cdots, t_{j-1}, [E1_{end}], \\ &\cdots, [E2_{start}], t_k, \cdots, t_{l-1}, [E2_{end}], \cdots, t_T], \end{aligned} \tag{1}$$

as the input token sequence for the *BERT* model.

To obtain a relation representation for two entities $E1$ and $E2$, we use the contextualized entity representation corresponding to the positions of $[E1_{start}]$ and $[E2_{start}]$ from *BERT*, rather than the $[CLS]$ token output that summarizes sentence-level semantics. These contextualized entity representations are then concatenated to form a fixed-length relation representation $\boldsymbol{m} \in \mathbb{R}^{2\times d}$, where d represents the hidden dimensions of 768.

Subsequently, we feed the representation $\boldsymbol{m}$ to the classifier head $fc(\cdot)$ to obtain the probability distribution $p(y|x) = fc(\boldsymbol{m})$ over the various classes and optimizing the *BERT* model with the cross-entropy loss function:

$$\mathcal{L}_x = \frac{1}{N_l} \sum_{i=1}^{N_l} loss(y_i, p(y|x_i)), \tag{2}$$

Similarly, we obtain a batch of probability distribution of unlabeled samples $\{p^i\}_{i=1}^{N_u}$ of size N_u using the fine-tuned *BERT* model. To visually depict the clustering of pseudo-labels, we employ a method for constructing pseudo-label graphs. The graph consists of nodes that correspond to samples and edges that indicate the similarity between pairs of samples. As a result, samples with higher similarity are located nearer to one another in the pseudo-label graph. We construct the pseudo-label graph via a similarity matrix W^p of size $N_u \times N_u$:

$$W_{ij}^p = \begin{cases} 1 & \text{if } i = j, \\ p_i \cdot p_j & \text{if } i \neq j \text{ and } p_i \cdot p_j \geq T, \\ 0 & \text{otherwise .} \end{cases} \tag{3}$$

For each pair of pseudo-labels with a similarity greater than the threshold T, a connection will be established, and each sample will be linked to itself with a self-loop of the strongest weight 1. The pseudo-label graph will serve as the target of the embedding graph in Representation Graph Learning to refine the representation space and obtain the RE task-specific representations.

3.2 Representation Graph Learning

Representation Graph Learning aims to acquire task-agnostic representations and utilize the available pseudo-labels to enhance task-specific representations for unlabeled data.

To generate the embedding graph, we apply a random transformation (e.g. synonym replacement) to each sentence except the entity part. As illustrated in Equation 1, we obtain a representation m' of the randomly transformed sentence. Then, we input $\boldsymbol{m'}$ and the representation m of the original sentence into the mapping head $map(\cdot)$ to generate low-dimensional embeddings $\boldsymbol{e} = map(\boldsymbol{m})$ and $\boldsymbol{e'} = map(\boldsymbol{m'})$ respectively. Inspired by Li et al. [14], we construct the embedding graph through the matrix W^e:

$$W_{ij}^e = \begin{cases} \exp\left(\boldsymbol{e}_i \cdot \boldsymbol{e}'_i/\tau\right) & \text{if } i = j, \\ \exp\left(\boldsymbol{e}_i \cdot \boldsymbol{e}_j/\tau\right) & \text{if } i \neq j. \end{cases} \tag{4}$$

We minimize the contrastive learning loss as self-ensembling loss function to optimize the embedding graph:

$$\mathcal{L}_u^{ctr} = -\frac{1}{N_u}\sum_{i=1}^{N_u} \log \frac{\exp\left(\boldsymbol{e}_i \cdot \boldsymbol{e}'_i/\tau\right)}{\sum_{j=1}^{N} \exp\left(\boldsymbol{e}_i \cdot \boldsymbol{e}_j/\tau\right)}, \tag{5}$$

where τ is the temperature coefficient. To acquire task-specific representations using the existing pseudo-labels, we will train the *BERT* model and the mapping head $map(\cdot)$ so that the embedding graph is analogous to the pseudo-label graph.

To make the two graphs comparable, we normalize the matrix W^p and W^e with $\widetilde{W}_{ij}^p = W_{ij}^p / \sum_j W_{ij}^p$ and $\widetilde{W}_{ij}^e = W_{ij}^e / \sum_j W_{ij}^e$ respectively. Then we could minimize the similarity loss of these two normalized graphs via:

$$\mathcal{L}_u^{sim} = \frac{1}{N_u}\sum_{i=1}^{N_u} \mathrm{H}\left(\widetilde{W}_i^p, \widetilde{W}_i^e\right), \tag{6}$$

where $\mathrm{H}(\widetilde{W}_i^p, \widetilde{W}_i^e)$ is defined as:

$$-\widetilde{W}_{ii}^p \log\left(\frac{\exp\left(\boldsymbol{e}_i \cdot \boldsymbol{e}'_i/\tau\right)}{\sum_{j=1}^{N_u} \widetilde{W}_{ij}^e}\right) - \sum_{j=1, j\neq i}^{N_u} \widetilde{W}_{ij}^p \log\left(\frac{\exp\left(\boldsymbol{e}_i \cdot \boldsymbol{e}_j/\tau\right)}{\sum_{j=1}^{N_u} \widetilde{W}_{ij}^e}\right). \tag{7}$$

The first term is a self-ensembling contrastive loss which motivates the model to produce embeddings that are alike for the original and transformed sentences. The second term pushes the model to group samples with comparable pseudo-labels to have similar embeddings. This clustering results in samples from the same class being positioned closer together, ensuring minimal entropy.

Our model could be self-refining with the pseudo label generation and representation graph learning during the training process. It will initially generate low-confidence pseudo-labels, resulting in a sparse pseudo-label graph. As the training progresses, the pseudo-label graph guides the embedding graph to enable the mapping head to generate task-specific representations, while the *BERT* model is optimized with the loss returned in this process, thus obtaining a more confident pseudo-label. The process of refining the model is iterative and continues until the stopping criterion is met, which in our case is 5 epochs, leading to a more accurate and reliable model. Our final loss function is:

$$\mathcal{L} = \mathcal{L}_x + \lambda_{ctr}\mathcal{L}_u^{ctr} + \lambda_{sim}\mathcal{L}_u^{sim}, \tag{8}$$

where the scalar hyper-parameters λ_{cls} and λ_{sim} are used to control the weight of the unsupervised losses.

4 EXPERIMENTS AND ANALYSES

4.1 Setup and Baselines

Datasets and Experimental Settings: Following previous works [9, 19], we evaluate the model on three public RE datasets: SemEval [7], which contains 6,507/1,493/2,717 data in train/dev/test sets and 19 relation types, with 17.4% no_relation. TACRED [29], which contains 68,124/22,631/15,509 data and 42 relation types, with 78.7% no_relation. Re-TACRED [23], which contains 58,465/19,584/13,418 data and 42 relation types, with 64.3% no_relation. We follow the existing setting [17] to use stratified sampling to divide the train set into various proportions of labeled and unlabeled sets to remains the same relation label distribution. Following previous works [9, 19], we sample 5%, 10%, and 30% of the training set as labeled data for the SemEval datasets, and 3%, 10%, and 15% of the training set as labeled data for TACRED and Re-TACRED datasets. For all datasets, we sample 50% of the train set as the unlabeled set. We adopt F1 Score as the evaluation metric. We use the BERT-Base default tokenizer with a max-length of 128 to preprocess data. For the classifier, we set the layer dimensions as 2×768-384-labels. For the projection head, we use a 2-layer MLP, we set the layer dimensions as 2×768-384-64. We use BertAdam [12] with a 3e-5 learning rate, warmup with 0.06 to optimize the loss, and set the batch size as 16. We set the temperature coefficient τ in Representation Graph Learning as 0.07. The hyper-parameters λ_{ctr} and λ_{sim} are set to 0.75 and 1.

Baselines: For baselines, we compare SelfLRE with nine competitive methods: (1) Self-Training [21], (2) Mean-Teacher [24], (3) DualRE [17], (4) RE-Ensemble [17], (5) MRefG [15], (6) MetaSRE [8], (7) GradLRE [9], (8) MixRE [25], and (9) UG-MCT [19]. These baselines belong to the self-training and self-ensembling methods. Finally, we present the upper bound model: BERT w. gold labels, which indicates that all unlabeled data have their gold labels during training with labeled data.

4.2 Results and Analysis

Main Results. Table 1 displays the F1 mean and deviation over 5 SemEval, TACRED, and Re-TACRED train/test runs, using different labeled data amounts and 50% unlabeled. We note that unlabeled data use enhances LRE models' performance compared to labeled-only data (BERT), showing unlabeled data integration improves RE task accuracy. SelfLRE consistently surpasses previous SOTA models MixRE and UG-MCT, with a 1.81% average improvement. Notably, when labeled data is scarce (e.g., 3% TACRED and Re-TACRED), SelfLRE achieves larger F1 improvement than the baselines. For instance, it registers a 4.01% improvement on 3% training set versus a 0.62% improvement on 15% set. We credit this to the self-refining framework, leveraging pseudo-labels for task-agnostic to RE task-specific representation learning, thereby iteratively securing better

Table 1: F1 (%) comparisons on the SemEval, TACRED and Re-TACRED datasets with various amounts of labeled data and 50% unlabeled data. The base encoders of all baselines are replaced by BERT for a fair comparison.

Methods / % Labeled Data	SemEval			TACRED			Re-TACRED		
	5%	10%	30%	3%	10%	15%	3%	10%	15%
BERT (Only labeled data)	70.71±1.24	71.93±0.99	78.55±0.87	40.11±3.88	53.17±1.67	55.55±0.82	42.64±1.24	58.45±1.38	64.34±1.02
Self-Training [21]	71.34±1.68	74.25±1.10	81.71±0.79	42.11± 1.04	54.17±0.53	56.52±0.40	46.32±0.87	62.65±0.75	66.42±0.98
Mean-Teacher [24]	70.05±3.89	73.37±1.42	80.61±0.81	44.34±1.78	53.08±1.01	53.79±1.38	45.64±1.32	61.32±0.83	66.64±1.35
RE-Ensemble [17]	72.35±2.63	75.71±1.39	81.34±0.74	42.78±1.89	54.83±0.95	55.68±1.21	46.84±2.33	64.23±1.34	67.42±1.05
DualRE-Pairwise [17]	74.35±1.76	77.13±1.10	82.88±0.67	43.06±1.73	56.03±0.55	57.99±0.67	48.95±1.59	65.39±1.21	68.21±0.86
DualRE-Pointwise [17]	74.02±1.68	77.11±1.02	82.91±0.62	43.73±1.60	56.28±0.61	57.72±0.49	49.42±1.33	65.67±1.02	68.98±1.21
MRefG [15]	75.48±1.34	77.96±0.90	83.24±0.71	43.81±1.44	55.42±1.40	58.21±0.71	48.83±1.35	65.24±1.32	68.39±0.83
MetaSRE [8]	78.33±0.92	80.09±0.78	84.81±0.44	46.16±1.02	56.95±0.34	58.94±0.36	54.34±2.32	67.83±1.45	70.24±1.73
GradLRE [9]	79.65±0.68	81.69±0.57	85.52±0.34	47.37±0.74	58.20±0.33	59.93±0.31	61.22±0.58	74.03±1.74	76.32±0.67
MixRE [25]	77.58±0.59	81.13±0.82	85.51±0.38	49.35±1.25	59.13±0.87	61.97±1.32	62.48±0.67	72.45±0.73	78.32±0.59
UG-MCT [19]	80.43±0.52	82.91±0.43	85.99±0.31	45.10±1.36	57.97±0.41	61.33±0.28	67.21±0.83	73.43±1.25	78.84±0.73
SelfLRE	**81.24±0.53**	**83.42±0.49**	**86.35±0.47**	**51.16±1.39**	**60.06±1.44**	**62.39±0.41**	**68.93±0.84**	**74.24±0.78**	**79.07±0.51**
w/o contrastive learning loss	77.23±0.74	80.55±0.62	84.19±0.47	49.68±1.31	58.41±1.22	61.32±0.95	66.43±1.84	73.52±1.57	78.45±0.69
w/o graph-based similarity loss	75.38±1.42	79.49±1.13	83.04±1.05	47.26±1.53	57.34±1.35	60.08±1.21	64.24±1.18	72.88±1.02	77.93±0.95
BERT w. gold labels	84.64±0.28	85.40±0.34	87.08±0.23	62.93±0.41	63.66±0.23	64.69±0.29	77.64±0.37	82.12±0.23	82.97±0.29

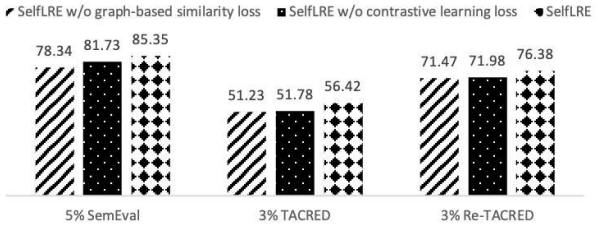

Figure 3: Pseudo label quality analysis on three datasets.

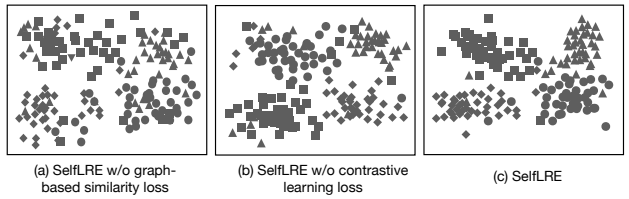

Figure 4: Visualizing contextualized representations after t-SNE dimension reduction. Features are shaped with their ground-truth relation labels.

pseudo labels. SelfLRE almost matches the performance of a model using gold labels, with only a 4.92% average difference across the datasets, as if using 50% more labeled data.

Ablation Study. We conduct an ablation study to showcase the impact of both loss functions in the test set. SelfLRE *w/o contrastive learning loss* means that the self-ensembling method is removed, which will weaken the model's ability to learn representations and affect the semantic analysis of the SelfLRE. SelfLRE *w/o graph-based similarity loss* means to remove the self-training method, leading to the unavailability of guidance from pseudo-labels in representation learning, which in turn affects task-specific representation learning. From the ablation rows in Table 1, we could observe that two loss functions all contribute positively to SelfLRE. Compared with contrastive learning loss, graph-based similarity loss can bring more performance improvement (3.24% vs. 1.89%), which shows the importance of pseudo-label supervision guidance.

Pseudo label Quality Analysis. We evaluate the contribution of the two modules to model performance by analyzing the F1 of the pseudo labels. As shown in Figure 3, we observe that both self-ensembling learning and self-training learning positively affect the model's performance. Among them, using high-quality pseudo-label data to guide task-specific representation learning can bring about a 5.71% F1 improvement, and the improved pseudo-label reversely promotes a more high-quality mapping network, resulting in further improvement of the pseudo-label classification ability.

Visualize Contextualized Representations. To demonstrate the impact of self-ensembling and self-training on relational representation learning, we used t-SNE [26] to visualize dimension-reduced representations. We selected 4 relations and 40 entity pairs from Re-TACRED and show the results in Figure 4. The SelfLRE *w/o graph-based similarity loss* already assigns meaningful semantics, but is inadequate for the RE task. Without contrastive learning loss, the model cannot provide confident clusters due to suboptimal learning. SelfLRE leverages the self-refining training schema to improve the relational representation learning — we could learn denser clusters and more discriminative representations.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel self-refining representation learning framework: SelfLRE for LRE task, which integrates two complementary self-training and self-ensembling methods. The self-training method could provide pseudo labels to help self-ensembling method refine task-specific representations. Conversely, the refined representations can be used to optimize pseudo-label classification to obtain higher-quality labels. Experiments on three datasets show that SelfLRE achieves 1.81% performance gain over SOTA baseline. In future work, we plan to extend the general LRE framework to other classification tasks, such as sentiment analysis, text classification, and also explore its applicability to other domains such as medical health and natural science.

ACKNOWLEDGMENTS

Lijie Wen is the corresponding author. Xuming Hu and Junzhe Chen have made equal contributions to this work. The work was supported by the National Key Research and Development Program of China (No. 2019YFB1704003), the National Nature Science Foundation of China (No. 62021002), NSF under grants III1909323, Tsinghua BNRist and Beijing Key Laboratory of Industrial Bigdata System and Application.

REFERENCES

[1] Yamen Ajjour, Pavel Braslavski, Alexander Bondarenko, and Benno Stein. 2022. Identifying argumentative questions in web search logs. In *Proc. of SIGIR*. 2393–2399.

[2] Soumen Chakrabarti. 2022. Deep knowledge graph representation learning for completion, alignment, and question answering. In *Proc. of SIGIR*. 3451–3454.

[3] Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Relation Extraction as Open-book Examination: Retrieval-enhanced Prompt Tuning. In *Proc. of SIGIR*. 2443–2448.

[4] Xiaokai Chu, Jiashu Zhao, Lixin Zou, and Dawei Yin. 2022. H-ERNIE: A multi-granularity pre-trained language model for web search. In *Proc. of SIGIR*. 1478–1489.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT*. 4171–4186.

[6] Asish Ghoshal, Srinivasan Iyer, Bhargavi Paranjape, Kushal Lakhotia, Scott Wen-tau Yih, and Yashar Mehdad. 2022. QUASER: Question Answering with Scalable Extractive Rationalization. In *Proc. of SIGIR*. 1208–1218.

[7] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of SemEval*. 33–38.

[8] Xuming Hu, Fukun Ma, Chenyao Liu, Chenwei Zhang, Lijie Wen, and Philip S Yu. 2021. Semi-supervised Relation Extraction via Incremental Meta Self-Training. In *Proc. of EMNLP: Findings*.

[9] Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and S Yu Philip. 2021. Gradient Imitation Reinforcement Learning for Low Resource Relation Extraction. In *Proc. of EMNLP*. 2737–2746.

[10] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[11] Gabriella Kazai, Paul Thomas, and Nick Craswell. 2019. The emotion profile of web search. In *Proc. of SIGIR*. 1097–1100.

[12] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

[13] Pranjal Kumar, Piyush Rawat, and Siddhartha Chauhan. 2022. Contrastive self-supervised learning: review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval* 11, 4 (2022), 461–488.

[14] Junnan Li, Caiming Xiong, and Steven CH Hoi. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proc. of ICCV*. 9475–9484.

[15] Wanli Li and Tieyun Qian. 2020. Exploit Multiple Reference Graphs for Semi-supervised Relation Extraction. *arXiv preprint arXiv:2010.11383* (2020).

[16] Wanli Li, Tieyun Qian, Xu Chen, Kejian Tang, Shaohui Zhan, and Tao Zhan. 2021. Exploit a Multi-head Reference Graph for Semi-supervised Relation Extraction. In *Proc. of IJCNN*. IEEE, 1–7.

[17] Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. In *Proc. of WWW*. 1073–1083.

[18] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (2021), 857–876.

[19] Bing Mao, Chang Jia, Yucheng Huang, Kai He, Jialun Wu, Tieliang Gong, and Chen Li. 2022. Uncertainty-guided Mutual Consistency Training for Semi-supervised Biomedical Relation Extraction. In *Proc. of BIBM*. IEEE, 2318–2325.

[20] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proc. of SIGIR*. 539–548.

[21] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. (2005).

[22] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proc. of ACL*. 2895–2905.

[23] George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proc. of AAAI*, Vol. 35. 13843–13850.

[24] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*. 1195–1204.

[25] Komal Kumar Teru. 2022. On Data Augmentation and Consistency-based Semi-supervised Relation Extraction. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*.

[26] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms. *The journal of machine learning research* 15, 1 (2014), 3221–3245.

[27] Junjie Yu, Xing Wang, Jiangjiang Zhao, Chunjie Yang, and Wenliang Chen. 2022. STAD: Self-Training with Ambiguous Data for Low-Resource Relation Extraction. In *Proc. of COLING*. 2044–2054.

[28] Shun Zhang, Xiangkui Lu, and Jun Wu. 2022. Co-Training with Validation: A Generic Framework for Semi-Supervised Relation Extraction. In *Proc. of CIKM*. 4697–4701.

[29] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proc. of EMNLP*. 35–45.